Deep Dive on Current and Future Flash Technology

Julie Herd

Director, Storage Product Management

Session 58677-2

DATA
FABRIC
NOW

Confidentiality Notice

The information in this presentation is confidential and proprietary to NetApp and may not be disclosed without the written permission of NetApp.

The information is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. NetApp makes no warranties, expressed or implied, on future functionality and timeline. The development, release, and timing of any features or functionality described for NetApp’s products remains at the sole discretion of NetApp. NetApp's strategy and possible future developments, products and or platforms directions and functionality are all subject to change without notice. NetApp has no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein.

Agenda

1. NAND Fundamentals and Trends
2. SSD Endurance
3. Emerging NAND Technologies

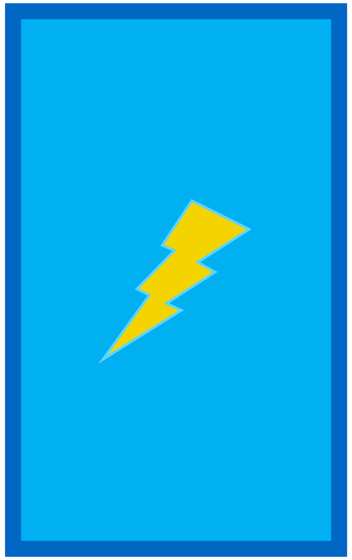
Fundamentals of Solid State

- SSDs are closer to memory than hard drives in terms of performance and endurance.
  - Much higher performance than HDD
  - Metrics for endurance are very different, new terms apply
- Solid-state technology is changing at a much faster rate than HDD
  - New capacities yearly, following the changes in wafer density
  - Multiple new technologies in development across NAND and storage-class memory
- Key takeaways from this session:
  - Understand the terms which define the performance and reliability of SSDs
  - Understand the standard endurance metrics which can be used to compare SSDs across vendors
  - Know how this applies to your deployment

NAND Fundamentals and Trends

NAND Types and the Shift to 3D NAND

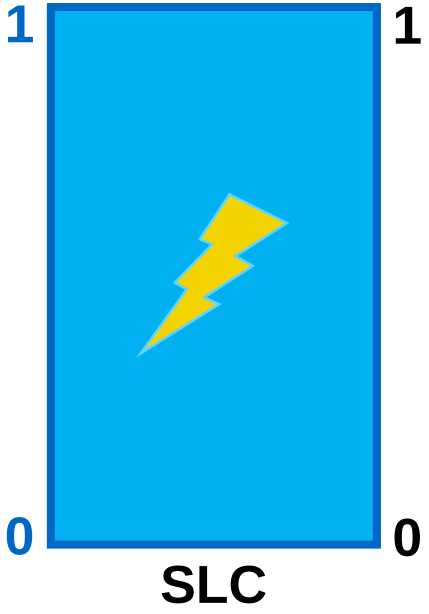
Basics of a NAND Cell



- A NAND cell is:
  - A container which stores a charge (electrons)
  - Surrounded by a permeable barrier to hold the charge
- A charge is stored in the cell
  - A charge is applied to insert electrons into the cell
  - The voltage amount determines the cell value
- NAND is non-volatile memory
  - Data is maintained across power cycles
  - The barrier requires power to release its charge

Evolution of NAND Density

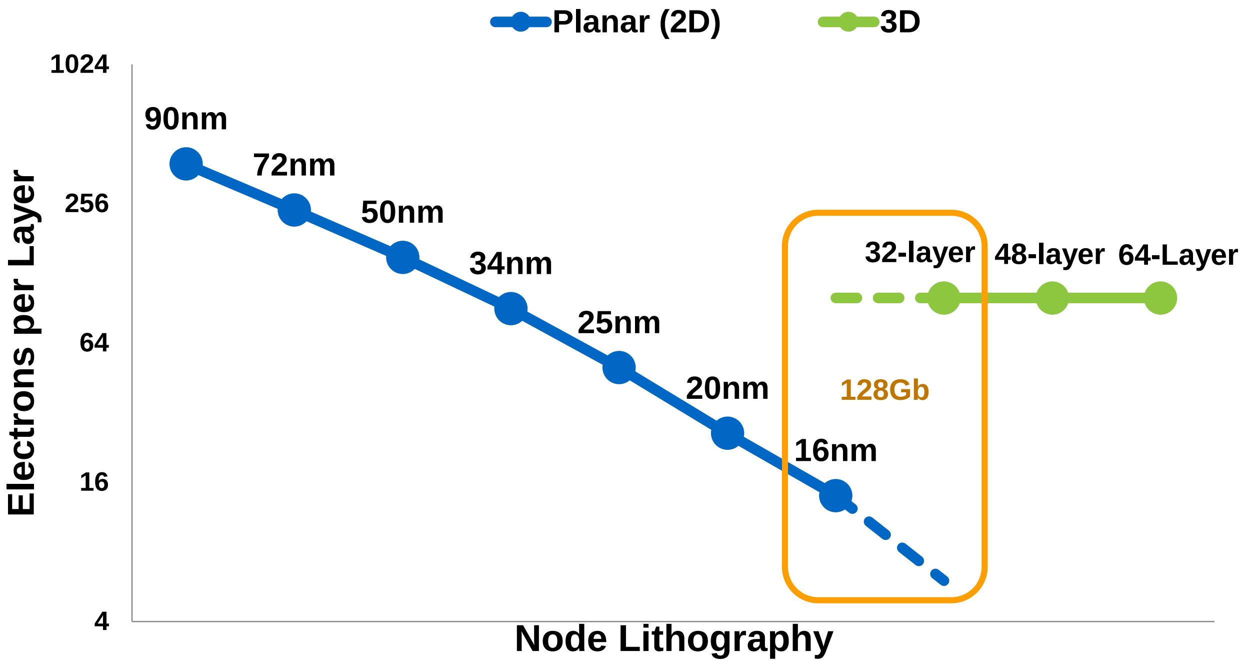
Increasing Densities of NAND Cells



Shift from Planar (2D) to 3D NAND

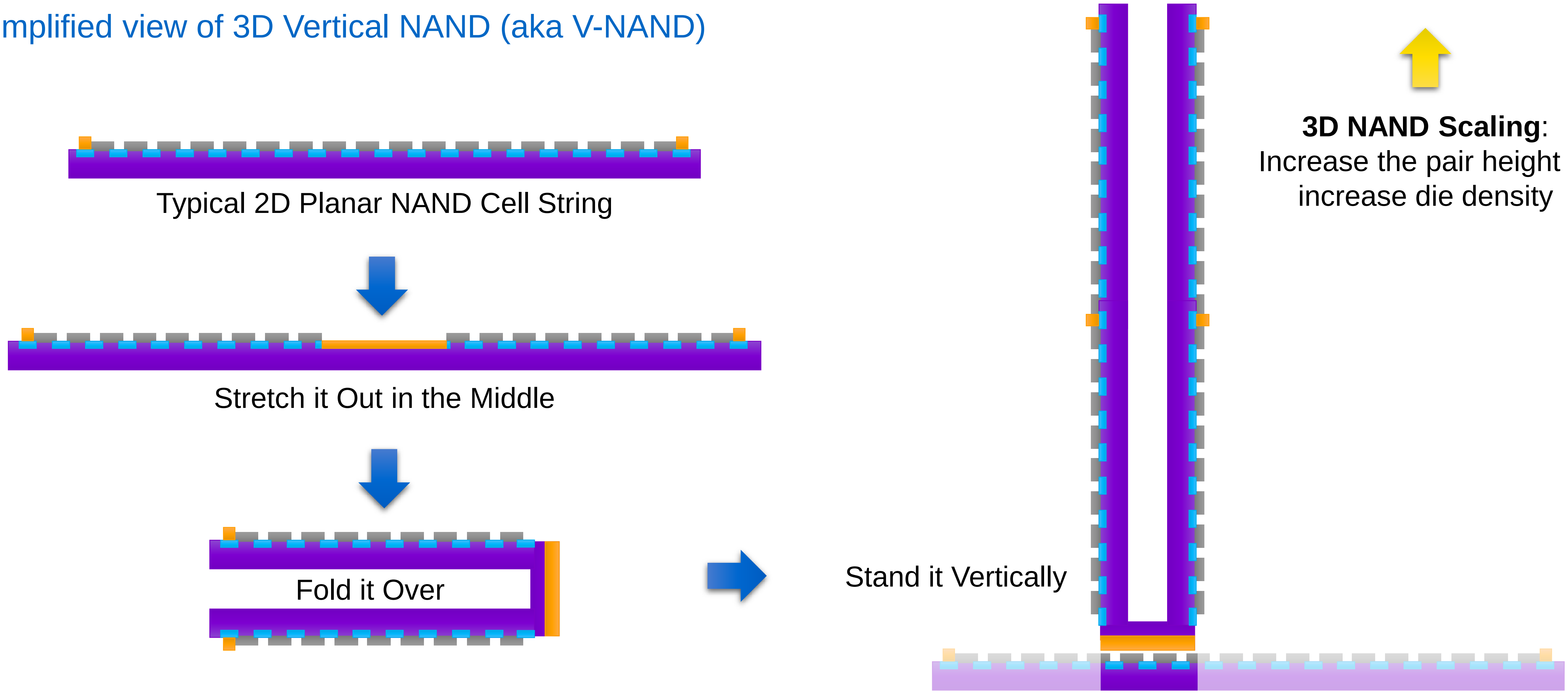
Scalability with reliability

- Shrinking planar cell holds less charge
  - Fewer electrons to distinguish between “0” and “1”
  - Small loss / gain of electrons can cause a bit flip
  - Fewer program / erase (P/E) cycles supported
- 3D NAND uses larger cells
  - More differentiation between bit states
  - Less interference between cells
  - Improved P/E support over planar NAND
- Increased scalability with 3D NAND
  - Planar NAND density is limited to 128Gb / die
  - 3D NAND density starts at 128Gb / die



Getting from Planar NAND to 3D NAND

Simplified view of 3D Vertical NAND (aka V-NAND)



Typical 2D Planar NAND Cell String

Stretch it Out in the Middle

Fold it Over

Stand it Vertically

3D NAND Scaling: Increase the pair height to increase die density

3D NAND uses less wafer area than 2D for the same bit density

Key Benefits of 3D NAND

- Endurance
  - Larger cell geometry provides more differentiation between bit states for MLC / TLC
  - Program / erase (P/E) cycle improved over planar NAND (and will continue thru ~2020)
- Density
  - Larger die sizes, along with bigger pages and blocks increases, overall density
  - Adding layers increases scalability while maintaining reliability
  - Smaller capacities become a challenge as layers increase
    - Planar NAND (15 / 16nm) will remain for several years to meet smaller scalability needs
- Power
  - New cell barrier technology (charge trap versus floating gate) reduces power consumption for program/erase
- Performance
  - Much better write performance over planar NAND
  - Read access times generally comparable to planar NAND

NAND Outlook

Near-term view of NAND Transitions

- NAND volume is currently shifting from planar 2D NAND to 3D NAND
  - Foundries are shifting volume production to new 3D NAND through 2017
  - Current 2D-node lithography (16-15nm) will remain for a few more years, primarily for smaller capacities
- All major NAND foundries have announced 3D-NAND SSDs
  - Samsung continues to lead production shipments of 3D NAND
  - Remaining foundries are transitioning to 3D NAND through 2016 / 2017
- Density
  - 128Gb/die will be highest planar 2D die density
  - 3D NAND has transitioned to 256Gb/die, with projections of 1Tb/die in future generations
  - QLC (MLC 4-bit) on the planning horizon, although remains several years away for production
- NAND remains least expensive ($/GB) non-volatile memory for the foreseeable future

SSD Endurance

Fundamentals of Solid-State Reliability

SSD Endurance Metrics

Measuring the reliability of an SSD

NAND Type != Endurance

- NAND is evolving quickly – NAND type (SLC, MLC, TLC) is not an indicator of endurance
  - The transition to 3D NAND changes the landscape for how endurance is measured
  - NAND type becomes a measure of scalability and cost efficiency

Old View

Type	Endurance
SLC	Best
eMLC	Better
cMLC	Good

New View – Coming Soon

SSD Endurance Factors

Knowing which SSD to choose

- Many factors affect SSD endurance ratings: “Your mileage may vary...”

Warranty

P/E Cycle Rating

Workload

Over-Provisioning %

SSD Capacity

Write Amplification

Voltage

Total Bytes Written

Data Retention at End of Life

Choosing the Right SSD

Matching endurance to workload

How Do I Know What To Choose?

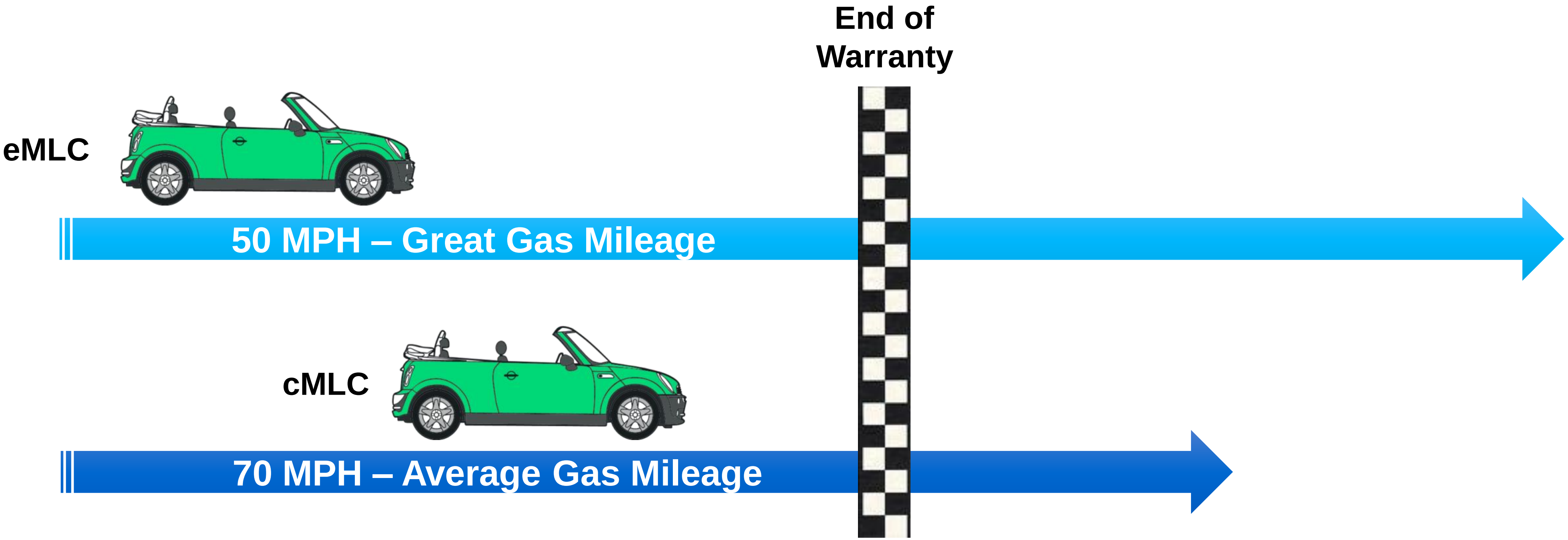
NetApp® Engineering does this work for you.

- Reliable performance and endurance for the full warranty period of the flash system
  - Tested across a variety of workloads and use cases
  - Extended warranty available, no caveats on usage
- Each flash system has unique requirements – SSDs are tailored to match
  - Validated against actual performance history (from AutoSupport)
  - Ensures all workloads are supported regardless of use case (Flash Pool™, hybrid, All Flash FAS, EF-Series, SolidFire®, etc)

SSD Endurance Fundamentals

- Program / Erase (P/E) cycles
  - SSDs have a finite lifetime, dictated by the number of write operations (P/E cycles) NAND cells can endure
  - Once the P/E cycles have been exceeded, the SSD cells are subject to wear-out
    - The permeable barrier around the cell begins to break down, allowing electrons to “leak” out
    - Bit errors occur when the charge can no longer be reliably determined, even with ECC (error-correcting code)
- A note on voltage
  - Using higher voltages to program NAND cells improves I/O performance, but accelerates wear-out
  - By writing to the flash more gently, the technology can be made to last considerably longer
  - This is the primary difference between eMLC and cMLC
    - eMLC = 20,000 to 30,000 P/E cycles, but slower programming speeds
    - cMLC = 3,000 to 10,000 P/E cycles

An Example of Endurance



End of Warranty

eMLC

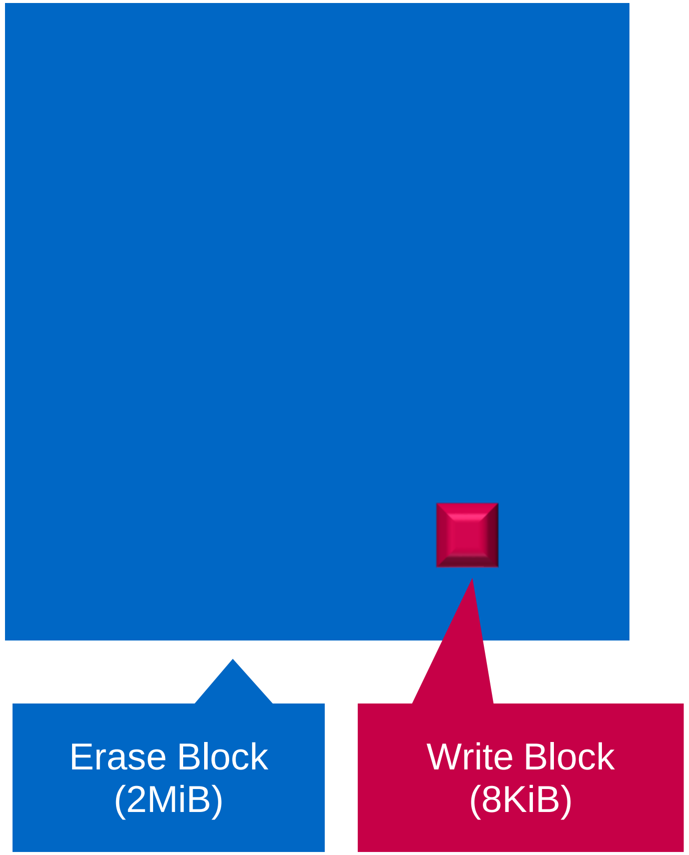
50 MPH – Great Gas Mileage

cMLC

70 MPH – Average Gas Mileage

Write Amplification

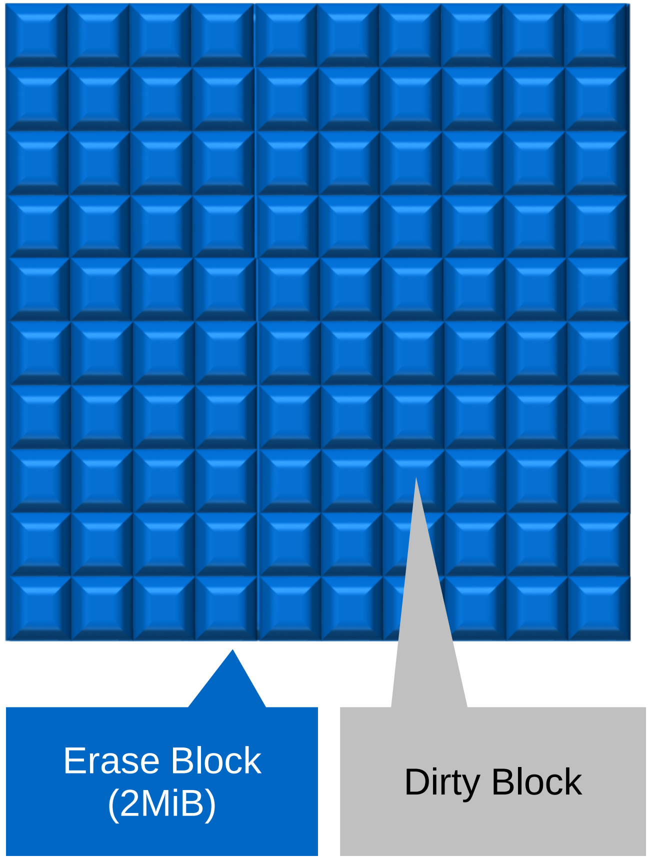
Unique characteristic of SSDs



- NAND flash must be erased before rewrite
  - Erase block is much larger than write block
  - Re-writing a block forces a rewrite of full erase unit
- Write amplification is
  - An aggregate measure of program / erase cycles
  - Caused by rewrite activity to a used block
- Write amplification example using fixed mapping
  - 4KiB block = 2MiB / 4KiB 512x WA
  - 32KiB block = 2MiB / 32KiB 64x WA

Reducing Write Amplification

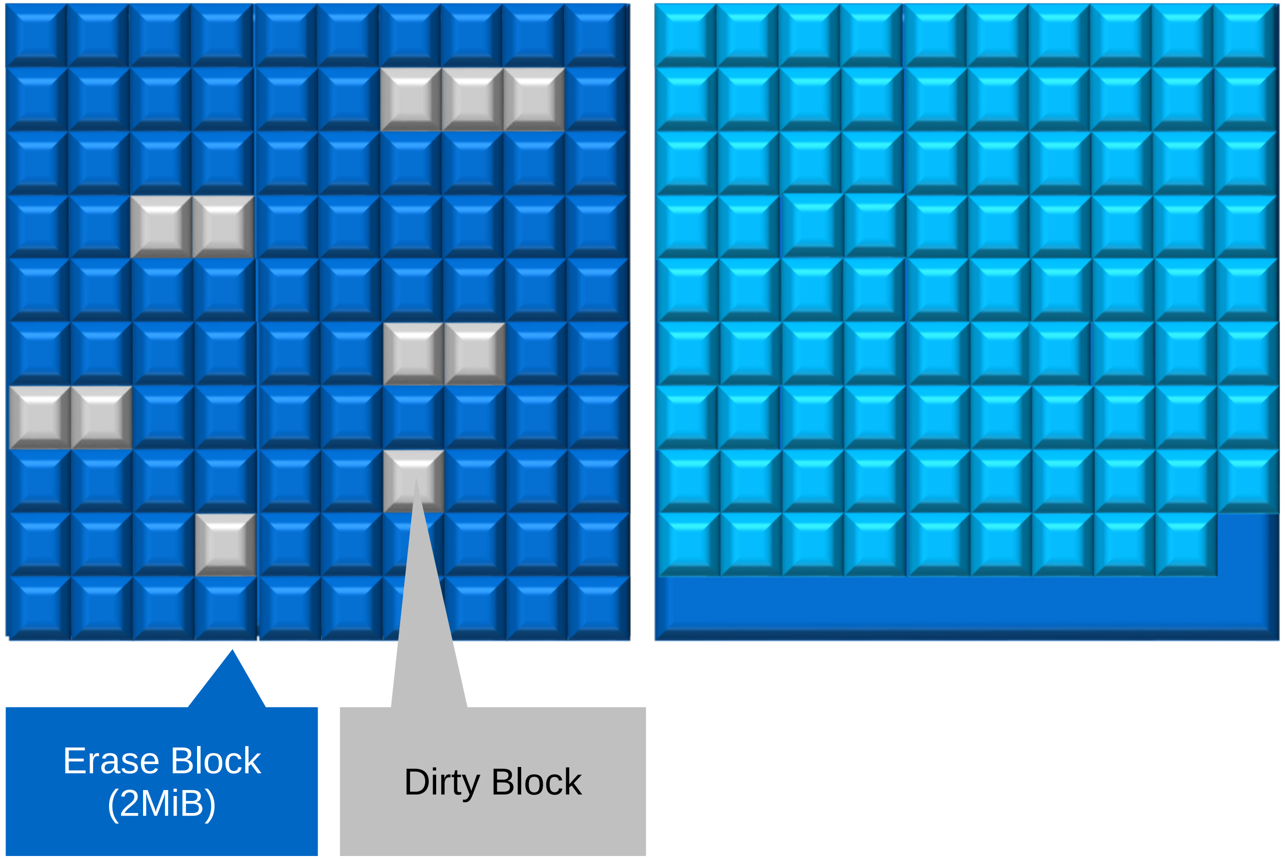
SSD Flash Translation Layer



- The Flash Translation Layer (FTL)
  - Abstraction layer between host address and on-disk data location
  - FTL can move on-disk data location as needed
- The FTL logs overwrite as they occur
  - New data is written to a new location
  - “Dirty” (not current) data is marked for deletion
  - Erased blocks are logged as dirty blocks as well
- Write amplification is reduced by deferring P/E cycle
  - Over time, block fragmentation occurs
  - Dirty blocks accumulate and need cleanup

Garbage Collection

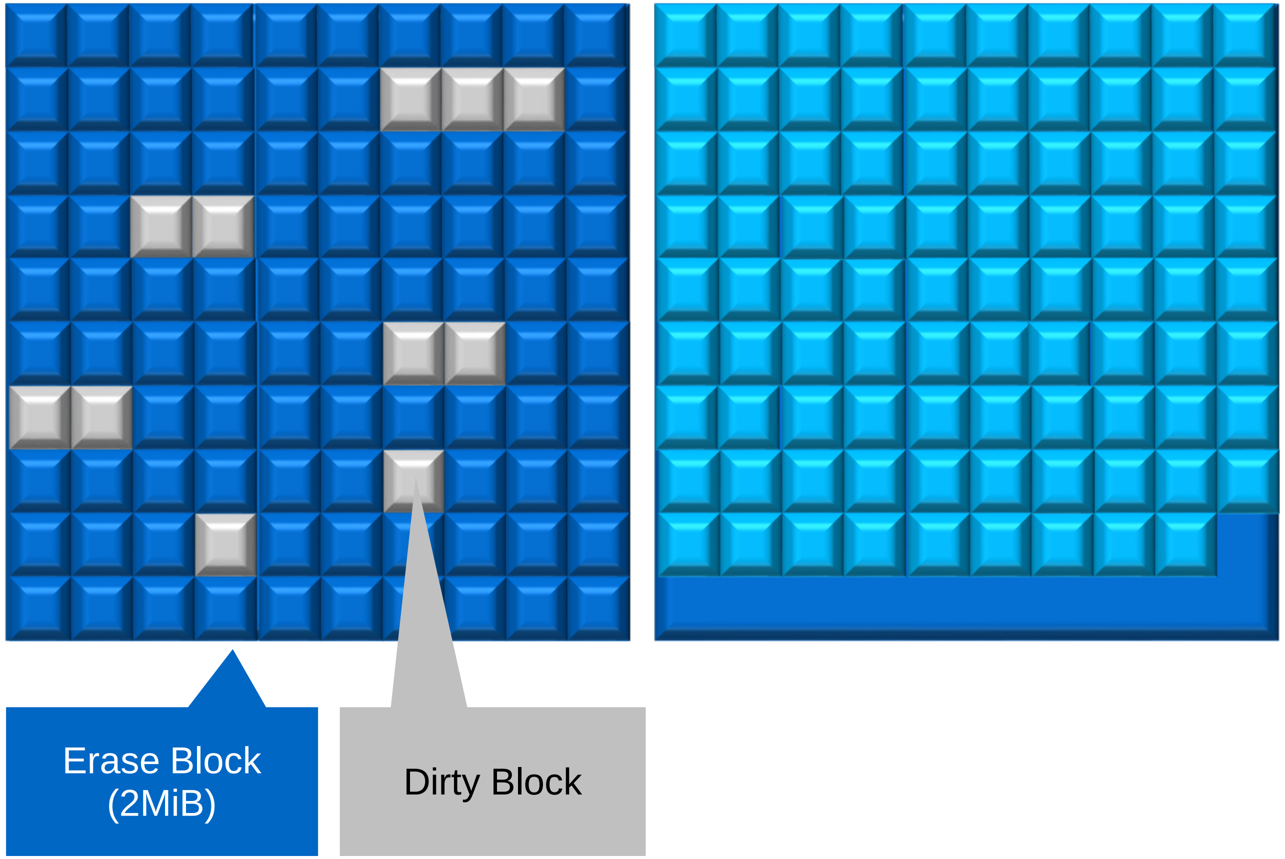
Reclaiming space



- The FTL can reclaim fragmented space
- To do that it must “garbage collect”
  - First read data blocks to be retained
  - Then write those blocks into an available erase block (free space on media)

Garbage Collection

Reclaiming space

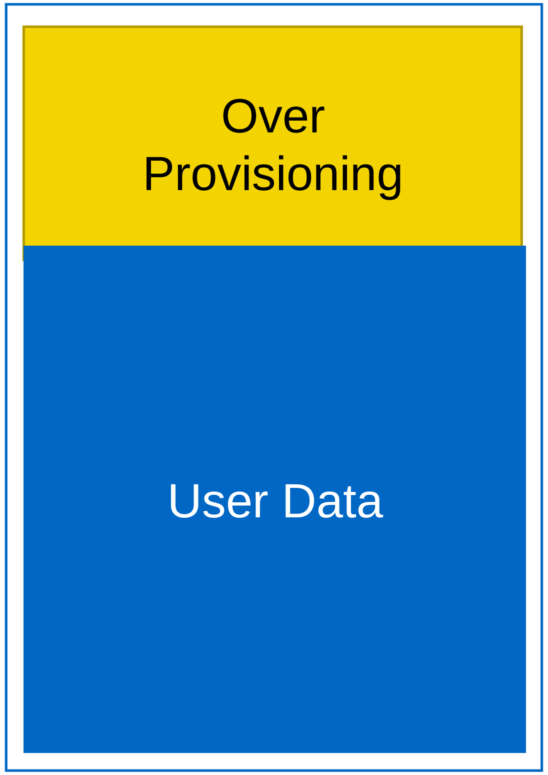


- The FTL can reclaim fragmented space
- To do that it must “garbage collect”
  - First read data blocks to be retained
  - Then write those blocks into an available erase block (free space on media)
  - Finally erase and prepare the erase block for new writes
- Garbage collection increases free space

Over Provisioning and Wear Leveling

Ensuring SSD performance and endurance

- Garbage collection needs “swap” space to operate efficiently
  - Over-provisioning (OP) guarantees that swap space is available
  - Reserved capacity on the SSD that can not be accessed by the user
  - Higher workloads need more OP space, read-intensive workloads need less
- The FTL will attempt to use all NAND erase blocks equally
  - Wear-leveling across all blocks to reduce the wear on individual cells
  - The OP area isn’t a fixed on-disk location, only a fixed capacity



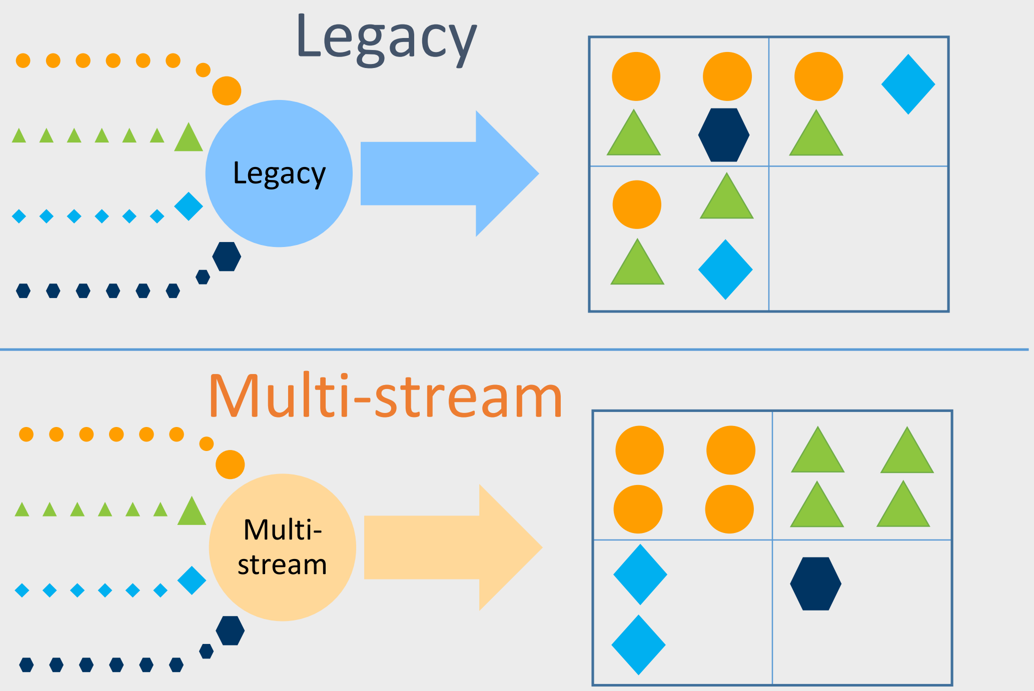
Multi-Stream Write

Reducing write amplification for SSDs

- Enables greater control and efficiency for data placement
  - Reduces the over-provisioning requirements by up to 20%
  - Enables more efficient data placement on the SSD by the operating system
  - Allows writes to be tagged, enabling the FTL to make better decisions over data placement
- T10 standard endorsed by all major SSD vendors
  - First vendor shipments begin in 2016
  - Full support by 2017

Data Placement Control

Highly efficient data placement inside SSD according to data characteristics (for example, controller, aggregate and other)

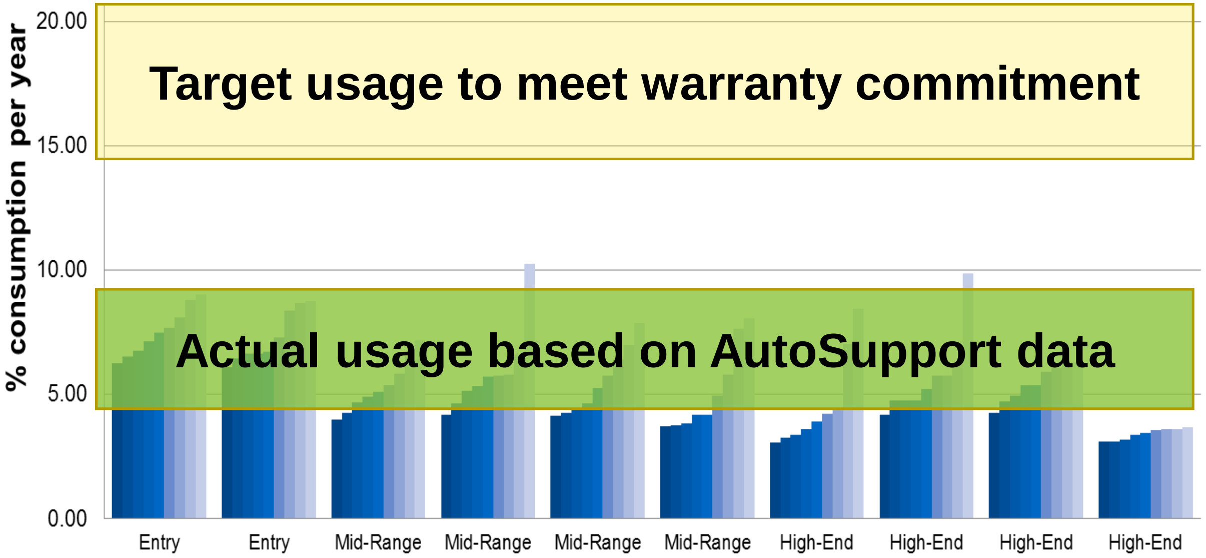


Total Bytes Written

NetApp AutoSupport (ASUP) Analysis

- Total bytes written (TBW)
  - Defines how much total data can be written to the SSD over the life of the drive
  - Based on
    - P/E cycle rating
    - Raw capacity
    - Expected write amplification

- NetApp® has a rich resource in ASUP for reporting real world SSD usage
  - SSDs report actual usage of "rated life used" across our installed base of SSD media
  - Reported as a % of TBW rating for the drive



SSD Endurance Metrics

Measuring the reliability of an SSD

- Drive writes per day (DWPD) is the primary metric for SSD endurance
  - How many times can the full SSD be written per day, for the full warranty period?
  - Used alongside of total bytes written to evaluate SSD endurance over the life of the device
  - A calculation based on
    - P/E cycle rating
    - Expected lifetime of the drive (aka warranty period)
    - Average write amplification (assuming a mixed read/write workload similar to SPC1)
    - Percentage of overprovisioned capacity

Old View

Type	Endurance
SLC	Best
eMLC	Better
cMLC	Good

New View

DWPD	Endurance Descriptions	
25	High Endurance	High End
10	Mainstream	High End
3	Value Endurance	Mid-Range
1	Read Intensive	
.5	Very Read Intensive	

MLC

TLC

DWPD — What Does It Really Mean?

Extreme use case example

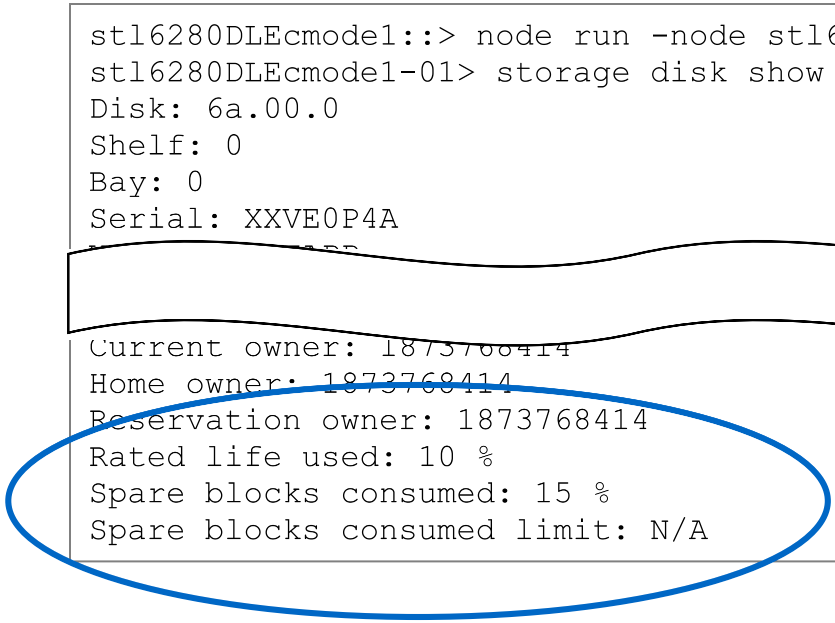
- All Flash FAS (AFF) A700 performance for 100% large, sequential writes is ~ 6500 MB/s
  - 100% writes every minute, all day would ingest 535 TB / day
  - Overwrite the same drives at max performance every day for 5 years
- How to handle this workload?

	960GB	**3.8T**	**15.3T**
# Drives	96	96	48
Raw Capacity	92.2TB	368.7TB	737.3TB
DWPD	3	1	1
% OP	7%	7%	7%
Throughput (TB / day)	**276.4**	**368.7**	**737.3**

Real-World SSD Usage

End-user reporting via ASUP

- AutoSupport field data reports key SSD usage metrics via <storage disk show> command
  - Rated life used – estimate based on
  - Spare blocks consumed - % of OP that has been consumed
  - Spare blocks consumed limit – threshold for trigger an ASUP event (in reality set at 90%)



```
stl6280DLEcmode1::> node run -node stl6280DLEcmode1-01
stl6280DLEcmode1-01> storage disk show -a 6a.00.0
Disk: 6a.00.0
Shelf: 0
Bay: 0
Serial: XXVE0P4A

Current owner: 1873768414
Home owner: 1873768414
Reservation owner: 1873768414
Rated life used: 10 %
Spare blocks consumed: 15 %
Spare blocks consumed limit: N/A
```

Emerging NAND Technologies

Changing landscape of Solid State

NVM Express (NMVe) Overview

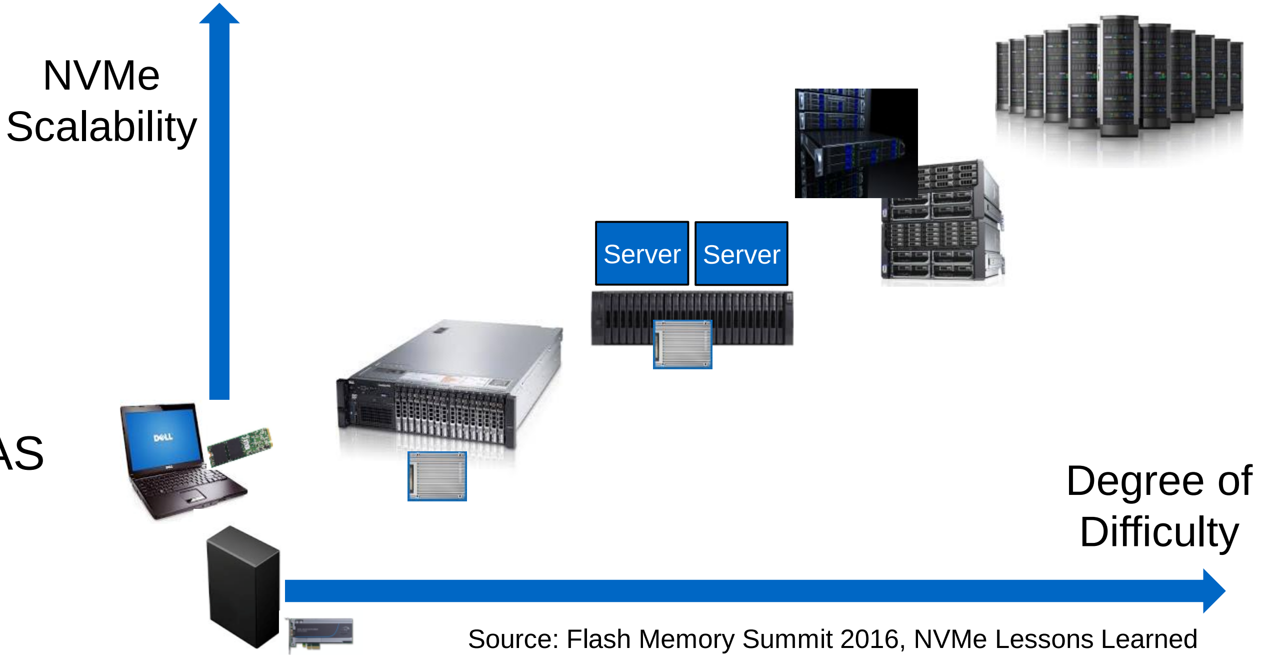
Next-Generation SSD Interface

- NVMe is a communication protocol developed specifically for SSDs
  - Utilizes high performance PCIe for connectivity to compute resources
  - Options for 2.5” SFF or M.2 (internal memory stick) form factors
- Benefits:
  - Efficiency = performance: low CPU utilization, low latency (streamlined interface)
  - Parallelism = performance; multiple cores, OS parallelism, I/O (64K queues, 64K queue depth)
- Challenges
  - PCIe does not natively hot-plug
    - Server / storage vendors have implemented solutions for “planned hot-plug”
    - “Surprise hot-plug” solutions are still under investigation
  - Lacks expansion capability outside the controller chassis
    - NVMe over fabric solutions underway to enable rack scale solutions

NVMe Adoption in the Industry

Server and storage adoption trends

- NVMe is overtaking SATA for internal SSD use
  - Internal cache
  - Data storage
  - NVMe $/GB will soon surpass SATA $/GB
- Server adoption underway
  - Single ported NVMe ideal for DAS
  - Strong performance for server uses
- Storage adoption on the horizon
  - Dual-ported NVMe needed to support HA pairs
  - Cost of dual-port NVMe not competitive with SAS

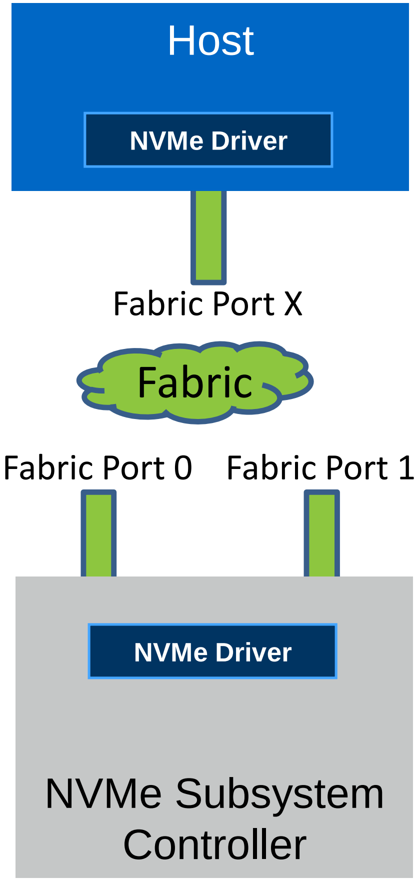


Source: Flash Memory Summit 2016, NVMe Lessons Learned

NVMe Over Fabrics

Expanding the NVMe Protocol beyond the chassis

- Extending NVMe over different fabric types
  - Maintains the benefits of NVMe across external connections
  - First definition is the RDMA protocol
    - Ethernet (iWARP and ROCE)
    - Infiniband
- Goals
  - Maintain end-to-end NVMe protocol, no translation
  - Equivalent IOPS performance between internal / external NVMe
  - Scalable to 100s of SSDs, well beyond PCIe-based attach



Dynamic SSD Capacity

Extending the usable lifetime of NAND SSDs

- As with HDDs, SSDs have capacity reserved for managing bad blocks within the SSD
  - Bad blocks could occur due to faulty dies, wear-out and more
  - The reserved block of capacity is not available for end use
  - If the number of bad blocks exceeds the reserved capacity, the SSD is failed

Actual NAND	0% OP	7% OP	28% OP
1024GB	1000GB	960GB	800GB
4096GB	4000GB	3840GB	3200GB

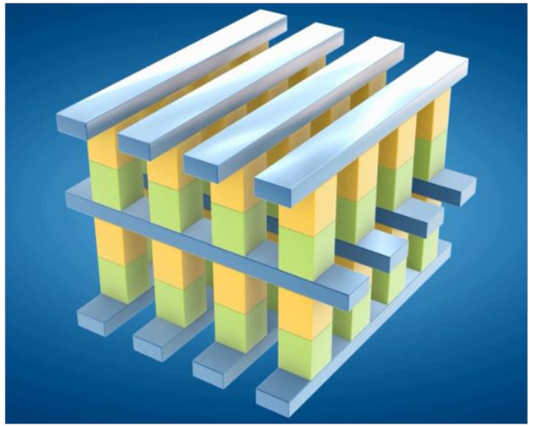
- Dynamic capacity would resize the SSD rather than fail due to excessive errors
  - Example: SSD to continue operations in the event of one or more bad dies on the SSD
  - Bad blocks would be marked and set as unused space by the SSD and the host
  - SSD usable space would be reset: for example, 800GB becomes 700GB

- Proposed standard – jointly developed by NetApp® and SSD vendors

Storage-Class Memories

Long-Term Alternatives to NAND

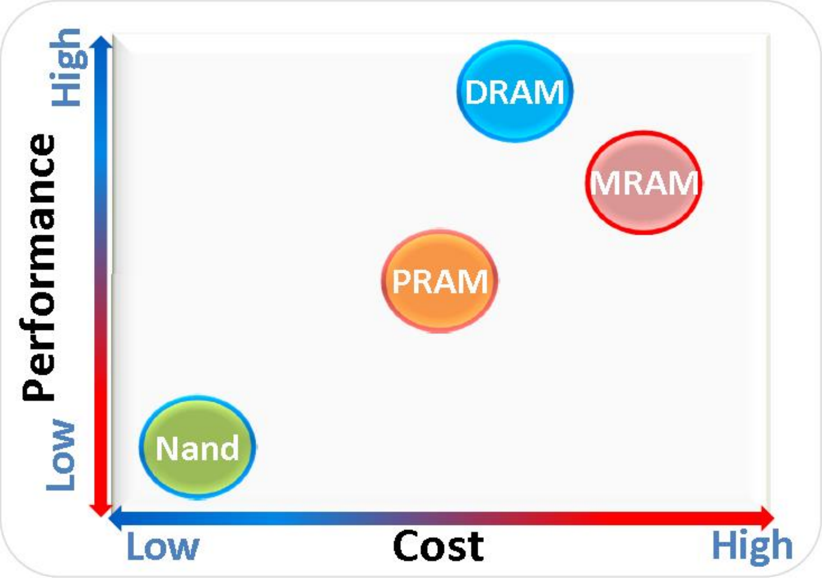
- Storage-class memories are “resistive” devices with long-term scalability
  - DRAM, NAND and HDDs are all electron devices, and ultimately face scaling challenges
- Multiple technologies are in development
  - 3DXpoint = Intel Optane (phase-change memory), can scale to additional layers and multiple bits per cell
  - STT-MRAM = spin-transfer torque magnetic RAM — closest to DRAM in speed and cost
  - ReRAM = Resistive RAM — huge potential in the future, not likely before 2020



Storage-Class Memory Comparison

	DRAM	MRAM	PRAM	NAND
Endurance	>10^{15}	>10^{15} *	~10^{7}	3 x 10^{3}
Operation Type	Byte	Byte	Byte	Page
Write Latency	≤ 15ns	≤ 45ns	~ 2µs *	~1.3ms
Read Latency	≤ 15ns	≤ 45ns	~110 ns *	~60µs
Retention	64ms	10 year *	10 years	10 years

* Target – still in development



Long-Term NAND Outlook

- NAND remains the dominant non-volatile memory technology through 2020
  - Price / performance are better suited for large-scale storage in near term
  - NAND remains the cheapest non-volatile memory ($/GB) for the foreseeable future

- Storage-class memory is an emerging market with high growth potential
  - Near-term replacement for DRAM for data retention and scalability
  - Current cost profile limits deployments to internal memory applications

Key Takeaways

- The transition to 3D NAND introduces new scalability with higher endurance.
- SSD endurance is dependent on many factors, including workload.
- NetApp® selects the right SSD for our system's workload profile.

Related Sessions and Resources

Learn More About this Topic

- 60616-2: Comparing Modern All-Flash Architectures
- 61692-2: All Flash FAS Technical Deep Dive
- 60615-3: NetApp® SolidFire® Technical Deep Dive

Tell Us What You Thought

Take an Insight survey on your mobile device.

Completed surveys are entered into a drawing to win prizes.

Don’t have the app? Download the Insight Mobile App
at the Apple Store and on Google Play.

Stay Connected

Follow NetApp Insight on Facebook and Twitter

Julie Herd

https://twitter.com/JulieHerd

Insight

https://www.facebook.com/NetAppInsight

https://twitter.com/NetAppInsight

#NetAppInsight

Thank you